

An Independent Process Approximation to Sparse Random Graphs with a Prescribed Number of Edges and Triangles

Stephen DeSalvo and M. Puck Rombach

September 30, 2015

Abstract

We prove a *pre-asymptotic* bound on the total variation distance between the uniform distribution over two types of undirected graphs with n nodes. One distribution places a prescribed number of k_T triangles and k_S edges not involved in a triangle independently and uniformly over all possibilities, and the other is the uniform distribution over simple graphs with exactly k_T triangles and k_S edges not involved in a triangle. As a corollary, for $k_S = o(n)$ and $k_T = o(n)$ as n tends to infinity, the total variation distance tends to 0, at a rate that is given explicitly. Our main tool is Chen–Stein Poisson approximation, hence our bounds are explicit for all finite values of the parameters.

Keywords. Poisson approximation, Random Graph Theory, Stein’s method, Asymptotic Enumeration of Graphs.

MSC classes: 05C30, 05A16, 05A20, 60C05

Many real-world networks display a property called *clustering* or *transitivity*, a dependency in the edge probability between two nodes on the number of common neighbours. In a human social network, for example, if two individuals have one or more friends in common, they are more likely themselves to be friends. Other types of networks, such as transportation networks, might display a negative transitivity, meaning that the probability of two nodes sharing a link decreases with the number of common neighbours, because such links may be unnecessary. The concept of clustering in social networks was introduced in [17]. Formally, the *clustering coefficient* C of a simple, undirected graph G is defined as

$$C(G) = \frac{\text{number of closed triplets of vertices in } G}{\text{number of connected triplets of vertices in } G},$$

where a closed triplet is an ordered triplet of vertices (i, j, k) such that the induced subgraph on (i, j, k) is a triangle, and a connected triplet is an ordered triplet of vertices (i, j, k) such that the induced subgraph on (i, j, k) is connected.

Many widely used models for real-world networks, such as random geometric graphs [9], random intersection graphs [5], random key graphs [18], the small-world model [17], and preferential attachment model [3, 7], naturally display some clustering, even though this is not an explicit parameter in their definitions. Regarding the latter, Bollobás and

Riordan show in [7] that it is possible to achieve almost any clustering coefficient or number of triangles by varying the parameters of the preferential attachment model as proposed in [3]. Several similar models with explicit clustering parameter have also been proposed [2, 12, 13, 15].

In [14], Newman proposes a uniform random graph model with specified number of edges and triangles for each vertex. This is a natural extension of the so-called *configuration model* [6], in which each vertex is given a specified number of half-edges, which are then joined up uniformly at random. In this paper, we make an attempt to study models where we control both the number of edges and the number of triangles in a graph, and sample from all such possible structures.

We will focus our attention on variations of the random graph model $\mathcal{G}(n, m)$. The more famous model $\mathcal{G}(n, p)$ is the most widely studied random graph model in probabilistic combinatorics. It was introduced by Gilbert [11] and developed by Erdős and Rényi to an extent that it is sometimes referred to as the Erdős-Rényi random graph model. Their work started with the introduction of a very similar model $\mathcal{G}(n, m)$ [10]. In the model $\mathcal{G}(n, p)$, there are n vertices and every edge (i, j) appears independently at random with probability $p = p(n)$. In $\mathcal{G}(n, m)$, there are n vertices and m edges, and the set of vertex pairs that have edges is chosen uniformly at random from all $\binom{n(n-1)/2}{m}$ such sets. When $m = np$, these two models behave similarly in the limit of $n \rightarrow \infty$ for many properties of interest, and $\mathcal{G}(n, p)$ is used more commonly because the independence between the edges allows for easier analysis. In this present work, however, it turns out that $\mathcal{G}(n, m)$ is much easier to analyze. See remark 1.3.

Our main result is a total variation distance bound between the distributions of the sets of edges in two different graph models, both of which are a variation on $\mathcal{G}(n, m)$. In the first model, we fix k_S edges and k_T triangles and place them *independently* at random. The resulting graph assumes no interaction between the edges and triangles, so one can imagine single edges as blue and the edges of each triangle as red, and we only count a triangle if it consists of all red edges, and similarly we do not count the red edges of a triangle as single blue edges; we denote by $\mathcal{M}(n, k_S, k_T)$ the random graph model which assigns equal probability to all such graphs. In the other model, we consider the set of all *simple* graphs on n nodes with exactly k_S edges not involved in a triangle and k_T triangles; we denote by $\mathcal{G}(n, k_S, k_T)$ the random graph model which assigns equal probability to all such graphs. Theorem 1.1 puts a maximum bound on the total variation distance between $\mathcal{M}(n, k_S, k_T)$ and $\mathcal{G}(n, k_S, k_T)$, which, asymptotically as the number of nodes n tends to infinity, tends to zero for $k_S = o(n)$ and $k_T = o(n)$.

To compute the total variation distance bound, we apply the Poisson approximation approach in [1]. Poisson approximation has a rich history, see [4] and the references therein. In particular, the application of Stein's method to the Poisson distribution [8], known as Chen-Stein Poisson approximation, not only allows one to prove Poisson convergence for certain collections of dependent random variables, it also provides a pre-asymptotic bound, *i.e.*, a bound which is explicit and absolute for all finite values of the parameters.

There are two dominant error terms in the approximation, which are the expected number of occurrences of clustering by three single edges to form an unintended triangle, and the expected number of occurrences of clustering by three triangles to form an extra, unintended triangle. We also use Theorem 1.1 to prove Proposition 3.1, which is a pre-

asymptotic estimate for the normalizing constant of $\mathcal{G}(n, k_S, k_T)$.

1 Random Graph Models

Let

1. $\mathcal{G}(n, m)$ denote graphs picked uniformly from the set of all simple graphs with exactly n nodes and m edges, with each of the $\binom{n}{2}$ graphs equally likely (the original Erdős–Rényi model [10]);
2. $\mathcal{M}(n, m)$ denote graphs picked uniformly from the set of all graphs with no self loops, but with multiple edges allowed, and with exactly n nodes and m edges, with each edge placed independently and uniformly at random from the $\binom{n}{2}$ possible edges, so that each of the $\frac{\binom{n}{2}^m}{m!}$ possible graphs are equally likely.

We now define the extensions of each of these sets in our context. Rather than just fix the number of edges, we fix the number of edges not involved in a triangle and also fix the number of triangles. Let

1. $\mathcal{G}(n, k_S, k_T)$ denote graphs picked uniformly from the set of all simple graphs with exactly n nodes, exactly k_S edges not part of any triangle, and exactly k_T triangles;
2. $\mathcal{M}(n, k_S, k_T)$ denote graphs formed by placing k_S edges independently and uniformly over all $\binom{n}{2}$ possible pairs of nodes, and placing k_T triangles independently and uniformly over all $\binom{n}{3}$ possible triplets of nodes on a graph. Again, multiple edges are allowed.

Let $\Omega_{\mathcal{M}}$ and $\mathbb{P}_{\mathcal{M}}$ denote the sample space and probability measure of $\mathcal{M}(n, k_S, k_T)$, respectively, and similarly let $\Omega_{\mathcal{G}}$ and $\mathbb{P}_{\mathcal{G}}$ denote the sample space and probability measure of $\mathcal{G}(n, k_S, k_T)$, respectively. Since $\Omega_{\mathcal{G}} \subset \Omega_{\mathcal{M}}$ and our measures are uniform, we define the set E to be such that

$$\mathbb{P}_{\mathcal{G}} = \mathbb{P}_{\mathcal{M}|E}, \tag{1}$$

i.e., the uniform distribution over elements in $\Omega_{\mathcal{G}}$. Thus our interest is in the quality of approximation of $\mathbb{P}_{\mathcal{M}|E}$ using $\mathbb{P}_{\mathcal{M}}$.

We now describe the set E in terms of random variables X_1, \dots, X_7 , where $X_i : \Omega_{\mathcal{M}} \rightarrow \mathbb{Z}$, $i = 1, \dots, 7$.

- $X_1 := \#\{\text{triplets of single edges that form a triangle}\},$
- $X_2 := \#\{\text{extra triangle by two single edges connecting to an edge of a triangle}\},$
- $X_3 := \#\{\text{extra triangle by single edge connecting two nodes in two triangles, touching in a third node}\},$
- $X_4 := \#\{\text{extra triangle formed by three intersecting triangles}\},$
- $X_5 := \#\{\text{double edge from a single edge on top of a triangle edge}\},$
- $X_6 := \#\{\text{double edges}\},$
- $X_7 := \#\{\text{double triangles}\}.$

Examples are shown in Figure 2. We have $E = \{X_1 = X_2 = \dots = X_7 = 0\}$.

A common measure of distance between two probability distributions is total variation distance. For any $k > 0$, given two \mathbb{R}^k -valued probability distributions $\mathcal{L}(X)$ and $\mathcal{L}(Y)$, the total variation distance between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ is denoted by $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y))$, and it is defined as

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{A \subseteq \mathbb{R}^k} |P(X \in A) - P(Y \in A)|, \quad (2)$$

where A is any Borel-measurable set.

We now present our main theorem, which is a quantitative bound on the total variation distance between $\mathbb{P}_{\mathcal{M}}$ and $\mathbb{P}_{\mathcal{G}}$

Theorem 1.1. *For all $n \geq 3$, $k_S \geq 0$, $k_T \geq 0$, we have*

$$(1 - e^{-\lambda}) - d_0 \leq d_{TV}(\mathbb{P}_{\mathcal{G}}, \mathbb{P}_{\mathcal{M}}) \leq (1 - e^{-\lambda}) + d_0, \quad (3)$$

where $\lambda = \sum_{i=1}^7 \lambda_i$ is given by Lemma 2.11 and d_0 is given by Lemma 2.12.

Note that Equation (3) is a *hard inequality*, i.e., not asymptotic. Now we specify the asymptotic range of parameter values for which Equation (3) tends to 0.

Corollary 1.2. *As n tends to infinity, we have*

$$d_0 = \Theta\left(\frac{(k_S + k_T)^4}{n^6}\right),$$

$$\lambda = \Theta\left(\frac{(k_S + k_T)^3 + k_T^2}{n^3} + \frac{k_T k_S + k_S^2}{n^2}\right).$$

Whence,

$$k_S = o(n) \text{ and } k_T = o(n) \iff d_{TV}(\mathbb{P}_{\mathcal{G}}, \mathbb{P}_{\mathcal{M}}) \sim \lambda \rightarrow 0.$$

Remark 1.3. Let $\mathcal{G}(n, p)$ denote the random graph model consisting of the set of graphs with exactly n nodes and each of the possible $\binom{n}{2}$ edges appearing independently with probability p , and also define its extension $\mathcal{G}(n, p_s, p_t)$, the set of graphs with exactly n nodes and each of the possible $\binom{n}{2}$ edges appearing independently with probability p_s and each of the possible $\binom{n}{3}$ triangles appearing independently with probability p_t . In this random graph model, the majority of the error in the approximation to $\mathcal{G}(n, k_S, k_T)$ is from the probability that the number of random edges/triangles is not *exactly* the values specified. This probability is given by a binomial distribution. Let S and T denote the number of edges and triangles, respectively, in a random graph generated by $\mathcal{G}(n, p)$ or $\mathcal{G}(n, p_s, p_t)$.

For $\mathcal{G}(n, p)$, with target k_S specified in advance, we optimally choose $p_s = k_S / \binom{n}{2}$, which yields

$$\mathbb{P}(S = k_S) = \binom{\binom{n}{2}}{k_S} p_s^{k_S} (1 - p_s)^{\binom{n}{2} - k_S} \sim \frac{1}{\sqrt{2\pi k_S (1 - k_S / \binom{n}{2})}}.$$

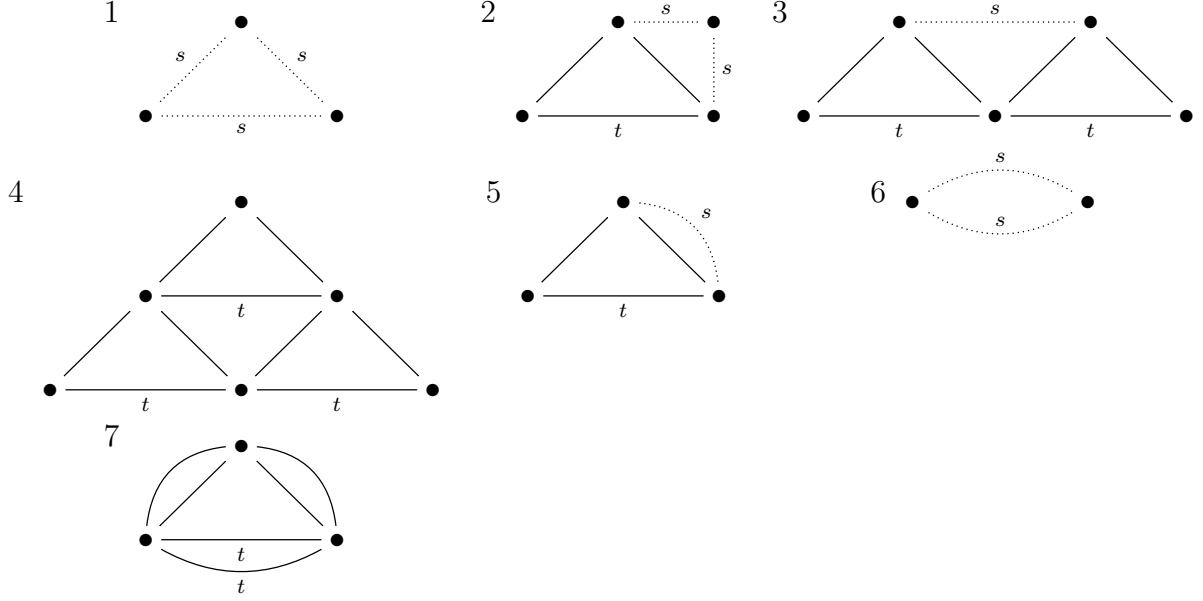


Figure 1: Pictorial representation of the bad events in a random graph model

Similarly, for $\mathcal{G}(n, p_s, p_t)$, with targets k_S and k_T specified in advance, we optimally choose $p_s = k_S / \binom{n}{2}$ and $p_t = k_T / \binom{n}{3}$, which by independence yields

$$\mathbb{P}(S = k_S, T = k_T) \sim \left(2\pi \sqrt{k_S k_T \left(1 - \frac{k_S}{\binom{n}{2}}\right) \left(1 - \frac{k_T}{\binom{n}{3}}\right)} \right)^{-1}.$$

Denote by \mathbb{P}_{p_s, p_t} the probability measure in the random graph model $\mathcal{G}(n, p_s, p_t)$. Letting $A = \{S = k_S, T = k_T\}$ in Equation (2), we have

$$d_{TV}(\mathbb{P}_G, \mathbb{P}_{p_s, p_t}) \geq 1 - P(S = k_S, T = k_T).$$

These calculations demonstrate that total variation distance is too strong of a metric to be used for this type of approximation.

2 Proof of Theorem 1.1

First, we note that as a consequence of Equation (1), and using $A = E$ in Equation (2), we have

$$d_{TV}(\mathbb{P}_G, \mathbb{P}_{\mathcal{M}}) = 1 - \mathbb{P}(X_1 = \dots = X_7 = 0) = 1 - \mathbb{P}(W = 0),$$

where $W = \sum_{i=1}^7 X_i$. Since the X_i , $i = 1, \dots, 7$ are not independent, we apply Chen–Stein Poisson approximation. Specifically, our proof is an application of Theorem 1 in [1].

We begin by defining the quantities b_1 and b_2 , which are required to specify the upper bound in Theorem 1.1. Suppose there is some countable or finite index set I . For each $\alpha \in I$, let Y_α denote an indicator random variable, with $p_\alpha := \mathbb{E} Y_\alpha = P(Y_\alpha = 1) > 0$, and $p_{\alpha\beta} := \mathbb{E} Y_\alpha Y_\beta$. Define $W := \sum_{\alpha \in I} Y_\alpha$, and $\lambda := \mathbb{E} W = \sum_{\alpha \in I} p_\alpha$. Next, for each

$\alpha \in I$, we define a dependency neighborhood B_α which consists of all indices $\beta \in I$ for which Y_α and Y_β are dependent. Then we define the quantities

$$b_1 := \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \quad (4)$$

$$b_2 := \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha\beta}, \quad (5)$$

$$b_3 := \sum_{\alpha \in I} \mathbb{E} |\mathbb{E} \{Y_\alpha - p_\alpha \mid \sigma(Y_\beta : \beta \notin B_\alpha)\}|. \quad (6)$$

Before one spends too much time parsing the precise meaning of b_3 , we note that when dependency neighborhoods B_α are chosen so that Y_α and Y_β are independent for $\beta \notin B_\alpha$, as in our setting, then $b_3 = 0$, and so it is just the first two quantities, b_1 and b_2 , which need to be computed.

Theorem 2.1 ([1]). *Let Z denote a Poisson random variable, independent of W , with $\mathbb{E} Z = \mathbb{E} W = \lambda$. Then we have*

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq 2(b_1 + b_2 + b_3),$$

and

$$|\mathbb{P}(W = 0) - e^{-\lambda}| \leq \min(1, \lambda^{-1})(b_1 + b_2 + b_3).$$

Denote the set of single edges by K_s , and the set of triangles by K_t . We define the unordered set of triplets of single edges,

$$\Gamma_1 := \{\{a, b, c\} : a, b, c \in K_s, \text{ distinct}\}.$$

Then $X_1 = \sum_{\alpha \in K_s} Y_\alpha$, where Y_α is the indicator random variable that the three single edges in α form a triangle.

Similarly, define

$$\Gamma_2 := \{\{a, b\}, \{\alpha\} : a, b \in K_s, \alpha \in K_t, \text{ distinct}\}.$$

Then, as before, $X_2 = \sum_{\alpha \in \Gamma_2} Y_\alpha$. The other cases for Γ_i and X_i , $i = 3, 4, 5, 6, 7$, are defined similarly.

Now we take $I = \bigcup_{i=1}^7 \Gamma_i$. Then for each $\alpha \in I$, we let B_α denote the set of indices β for which Y_α and Y_β are independent; this is precisely the set of collections of indices where α and β share any combination of *at least 2* single edges or triangles.

For example, when $\alpha = \{a, b, c\} \in \Gamma_1$, then B_α is the set of all Y_β , $\beta \in I$ which contain at least two of a, b , or c . For $\beta = \{\{a, b\}, \{u\}\} \in \Gamma_2$, we have $\beta \in B_\alpha$. We also have $Y_\beta \notin B_\alpha$ for any $\alpha \in \Gamma_1$ and $\beta \in \Gamma_4$, since Γ_1 only consists of single edges and Γ_4 only consists of triangles.

When $\alpha, \beta \in I$ do not share any elements, it is obvious that Y_α and Y_β are independent. Furthermore, even when α, β share exactly one element, Y_α and Y_β are still independent, since conditioning on an occurrence of Y_α does not give any information regarding *where* the occurrence occurred.

It now remains to compute the desired quantities; we start with b_1 . By simple counting arguments, we have the following.

Lemma 2.2.

$$p_\alpha = \left\{ \begin{array}{ll} \frac{\binom{n}{3}}{\binom{n}{2}^3}, & \alpha \in \Gamma_1, \\ \frac{\binom{n}{4}\binom{4}{3} \cdot 3}{\binom{n}{3}\binom{n}{2}^2}, & \alpha \in \Gamma_2, \\ \frac{\binom{n}{5}\binom{5}{3}\binom{3}{2}}{\binom{n}{3}^2\binom{n}{2}}, & \alpha \in \Gamma_3, \\ \frac{\binom{n}{6}\binom{6}{3}\binom{3}{2}^2}{\binom{n}{3}^3}, & \alpha \in \Gamma_4, \\ \frac{\binom{n}{3} \cdot 3}{\binom{n}{3}\binom{n}{2}}, & \alpha \in \Gamma_5, \\ \frac{\binom{n}{2}}{\binom{n}{2}^2}, & \alpha \in \Gamma_6, \\ \frac{\binom{n}{3}}{\binom{n}{2}}, & \alpha \in \Gamma_7. \end{array} \right\} \quad (7)$$

Next, since the terms p_α and p_β in the sum in Equation (4) do not depend on the particular set of indices, we simply need to count the number overlapping indices for which Y_α and Y_β are dependent; when $\alpha \in \Gamma_i$ and $\beta \in \Gamma_j$, we denote the number of indices $\beta \in B_\alpha$ by $C_{i,j}$. The lemma below follows by straightforward counting.

Lemma 2.3. *We have*

$$\begin{array}{llll} C_{1,1} = \binom{k_S}{4} \binom{4}{2}, & C_{1,2} = \binom{k_S}{3} \binom{3}{2} k_T, & C_{1,3} = 0, & C_{1,4} = 0 \\ C_{1,5} = 0 & C_{1,6} = \binom{k_S}{3} \binom{3}{2} & C_{1,7} = 0 & C_{2,2} = \binom{k_S}{2} \binom{k_T}{2} + \binom{k_S}{3} \binom{3}{1} k_T \\ C_{2,3} = \binom{2}{1} \binom{k_S}{2} \binom{2}{1} \binom{k_T}{2} & C_{2,4} = 0 & C_{2,5} = \binom{k_S}{2} \binom{2}{1} k_T & C_{2,6} = \binom{k_S}{2} k_T \\ C_{2,7} = 0 & C_{3,3} = \binom{k_T}{2} \binom{k_S}{2} + k_S \binom{k_T}{3} \binom{3}{1} & C_{3,4} = k_S \binom{k_T}{3} \binom{3}{1} & C_{3,5} = k_S \binom{k_T}{2} \binom{2}{1} \\ C_{3,6} = 0 & C_{3,7} = \binom{k_S}{1} \binom{k_T}{2} & C_{4,4} = \binom{k_T}{4} \binom{4}{2} & C_{4,5} = 0 \\ C_{4,6} = 0 & C_{4,7} = \binom{k_T}{3} \binom{3}{2} & C_{5,5} = 0 & C_{5,6} = 0 \\ C_{5,7} = 0 & C_{6,6} = 0 & C_{6,7} = 0 & C_{7,7} = 0. \end{array}$$

We can now state the formula for b_1 below.

Proposition 2.4. *For the random graph model $\mathcal{M}(n, k_s, k_t)$, we have*

$$b_1 = \sum_{i=1}^7 \sum_{j=i}^7 C_{i,j} p_i p_j, \quad (8)$$

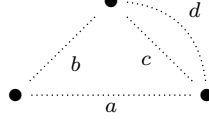
where the values are specified in Lemmas 2.2, 2.3, and 2.10.

Corollary 2.5. *Asymptotically, as $n \rightarrow \infty$, we have*

$$b_1 \sim O\left(\frac{(k_S + k_T)^4}{n^6}\right).$$

The expressions for $p_{\alpha\beta}$ are no more difficult to calculate, although care must be taken to account for all possible symmetries. Since the values for p_α , $\alpha \in I$ have already been specified, we instead focus on calculating $p_{\beta|\alpha} := \mathbb{E}(Y_\beta|Y_\alpha)$, since $p_{\alpha\beta} = p_\alpha p_{\beta|\alpha}$.

To demonstrate, we fix some $\{a, b, c\} = \alpha \in \Gamma_1$, then we consider $\beta \in \Gamma_1$, $\beta \neq \alpha$. There is only one case to consider; that is, when $\beta = \{a, b, d\}$. This scenario can be described pictorially as



In this case, Y_α and Y_β are dependent since knowing that edges a and b are already in a triangular formation affects the probability that a , b , and d are in a triangular formation. In fact, in this case we have $p_{\beta|\alpha} = \binom{n}{2}^{-1}$, since there is only one location allowed for edge d , i.e., it must coincide with edge c .

Also, the terms $C_{2,2}$ and $C_{3,3}$ are the sum of two distinct forms of overlapping, and each has a different corresponding conditional probability. Thus we subdivide these terms into $C_{2,2} = C_{2,2}^1 + C_{2,2}^2$ and $C_{3,3} = C_{3,3}^1 + C_{3,3}^2$.

Lemma 2.6. *Let $C_{2,2}^1$ denote the number of indices $\alpha, \beta \in \Gamma_2$ which do not share a triangle. Let $C_{2,2}^2$ denote the number of indices $\alpha, \beta \in \Gamma_2$ which do share a triangle. Let $C_{3,3}^1$ denote the number of indices $\alpha, \beta \in \Gamma_3$ which share both triangles. Let $C_{3,3}^2$ denote the number of indices $\alpha, \beta \in \Gamma_3$ which share a triangle and a single edge. Then we have*

$$\begin{aligned} C_{2,2}^1 &:= \binom{k_S}{2} \binom{k_T}{2}, & p_{2|2}^1 &= \binom{n}{3}^{-1}, \\ C_{2,2}^2 &:= \binom{k_S}{3} \binom{3}{1} k_T, & p_{2|2}^2 &= \binom{n}{2}^{-1}, \\ C_{3,3}^1 &:= \binom{k_T}{2} \binom{k_S}{2}, & p_{3|3}^1 &= \binom{n}{2}^{-1}, \\ C_{3,3}^2 &:= k_S \binom{k_T}{3} \binom{3}{1}, & p_{3|3}^2 &= \binom{n}{3}^{-1}. \end{aligned}$$

Lemma 2.7. *We have*

$$p_{1|1} = \binom{n}{2}^{-1},$$

$$p_{2|1} = p_{3|2} = p_{4|3} = p_{4|4} = \binom{n}{3}^{-1},$$

and the rest are either specified in Lemma 2.6 or are 0.

In order to more easily state the bound for b_2 , we make a final definition, which is for a collection of constants $C_{i,j}^*$, $i, j = 1, \dots, 7$, $i \leq j$.

Definition 2.8. *We define*

$$C_{2,2}^* := C_{2,2}^1 p_2 p_{2|2}^1 + C_{2,2}^2 p_2 p_{2|2}^2,$$

$$C_{3,3}^* := C_{3,3}^1 p_3 p_{3|3}^1 + C_{3,3}^2 p_3 p_{3|3}^2.$$

For $i \leq j$, $i, j = 1, \dots, 7$, excluding the cases $i = j = 2$ and $i = j = 3$, we define

$$C_{i,j}^* := C_{i,j} p_i p_{j|i}.$$

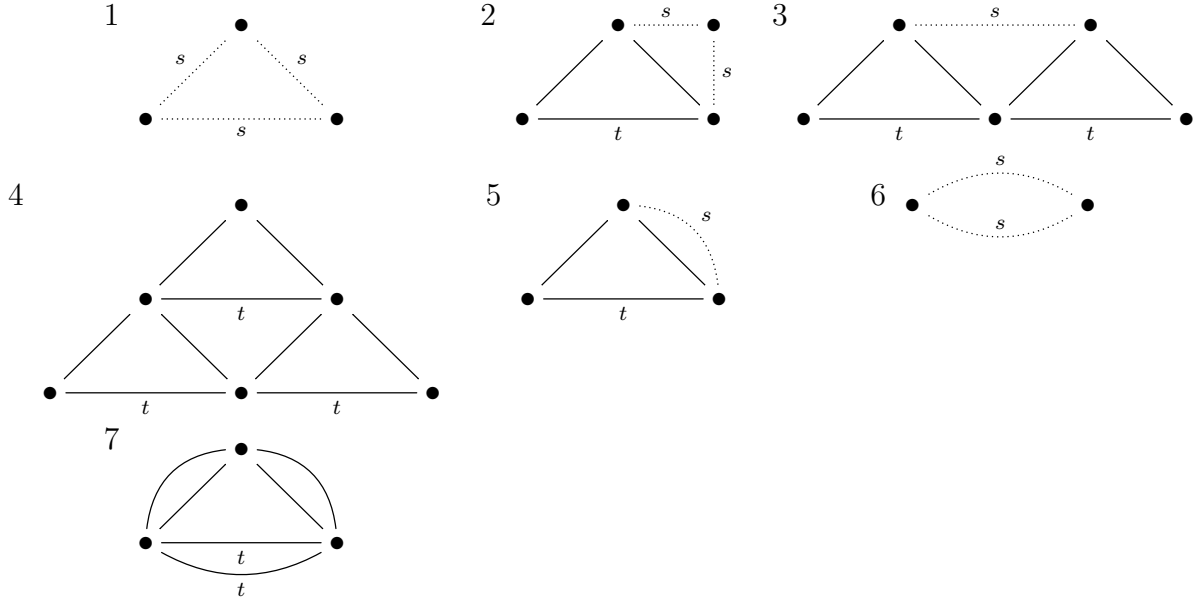


Figure 2: Pictorial representation of the bad events in a random graph model

Proposition 2.9. *We have*

$$b_2 = \sum_{i=1}^7 \sum_{j=i}^7 C_{i,j}^* \sim O\left(\frac{(k_S + k_T)^4}{n^6}\right).$$

Next, we calculate the sizes of each set Γ_i , $i = 1, \dots, 7$.

Lemma 2.10.

$$\begin{aligned} |\Gamma_1| &= \binom{k_s}{3}, \\ |\Gamma_2| &= \binom{k_s}{2} k_t, \\ |\Gamma_3| &= k_s \binom{k_t}{2}, \\ |\Gamma_4| &= \binom{k_t}{3}, \\ |\Gamma_5| &= k_s k_t, \\ |\Gamma_6| &= \binom{k_s}{2}, \\ |\Gamma_7| &= \binom{k_t}{2}. \end{aligned}$$

Lemma 2.11. *Let*

$$\lambda_j := \mathbb{E} X_j = |\Gamma_j| p_j, \quad j = 1, 2, \dots, 7.$$

In particular, we have

$$\begin{aligned}
\lambda_1 &= \frac{\binom{k_s}{3}\binom{n}{3}}{\binom{n}{3}^3} \sim \frac{2}{9} \frac{k_s^3}{n^3} = O\left(\frac{k_s^3}{n^3}\right) \\
\lambda_2 &= \frac{\binom{k_s}{2}k_t\binom{n}{4}\binom{4}{3} \cdot 3}{\binom{n}{3}\binom{n}{2}^2} \sim 6 \frac{k_s^2 k_t}{n^3} = O\left(\frac{k_s^2 k_t}{n^3}\right) \\
\lambda_3 &= \frac{\binom{k_t}{2}k_s\binom{n}{5}\binom{5}{3}\binom{3}{2}}{\binom{n}{3}^2\binom{n}{2}} \sim 18 \frac{k_t^2 k_s}{n^3} = O\left(\frac{k_t^2 k_s}{n^3}\right) \\
\lambda_4 &= \frac{\binom{k_t}{3}\binom{n}{6}\binom{6}{3}\binom{3}{2}^2}{\binom{n}{3}^3} \sim 9 \frac{k_t^3}{n^3} = O\left(\frac{k_t^3}{n^3}\right) \\
\lambda_5 &= \frac{k_t k_s \binom{n}{3} \cdot 3}{\binom{n}{3}\binom{n}{2}} \sim 6 \frac{k_t k_s}{n^2} = O\left(\frac{k_t k_s}{n^2}\right) \\
\lambda_6 &= \frac{\binom{k_s}{2}\binom{n}{2}}{\binom{n}{2}^2} \sim \frac{k_s^2}{n^2} = O\left(\frac{k_s^2}{n^2}\right) \\
\lambda_7 &= \frac{\binom{k_t}{2}\binom{n}{3}}{\binom{n}{3}^2} \sim 3 \frac{k_t^2}{n^3} = O\left(\frac{k_t^2}{n^3}\right).
\end{aligned}$$

Lemma 2.12. Let $W = \sum_{j=1}^7 X_j$, and define $\lambda := \mathbb{E}W$. Suppose Z is an independent Poisson random variable with expected value λ . Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq 2(b_1 + b_2),$$

where b_1 and b_2 are defined by Equation (4) and Equation (5), respectively. In addition, we have

$$(1 - e^{-\lambda}) - d_0 \leq P(W > 0) \leq (1 - e^{-\lambda}) + d_0, \quad (9)$$

where $d_0 = \min(1, \lambda^{-1})(b_1 + b_2)$.

3 Applications

3.1 The number of graphs in $\mathcal{G}(n, k_S, k_T)$

Let us start by comparing the two simpler graph models $\mathcal{G}(n, m)$ and $\mathcal{M}(n, m)$. We slightly abuse notation to let $|\mathcal{G}(n, m)|$ and $|\mathcal{M}(n, m)|$ denote the number of possible graphs in these respective probability spaces. We have

$$|\mathcal{G}(n, m)| = \binom{\binom{n}{2}}{m}, \quad |\mathcal{M}(n, m)| = \frac{\binom{n}{2}^m}{m!}.$$

Let \mathbb{P}_1 and \mathbb{P}_2 denote the probability measures of $\mathcal{G}(n, m)$ and $\mathcal{M}(n, m)$, respectively. There exists a coupling of \mathbb{P}_1 and \mathbb{P}_2 so that the random graph model either generates the same graph in both $\mathcal{G}(n, m)$ and $\mathcal{M}(n, m)$, or generates a graph in $\mathcal{M}(n, m) \setminus \mathcal{G}(n, m)$. The total variation distance in this case is simply

$$\begin{aligned}
d_{TV}(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{P}(\text{graph model generated in } \mathcal{M}(n, m) \setminus \mathcal{G}(n, m)) \\
&= 1 - \prod_{i=1}^{m-1} \left(1 - \frac{i}{\binom{n}{2}}\right) \sim 1 - e^{-m^2/n^2},
\end{aligned}$$

whence

$$d_{TV}(\mathbb{P}_1, \mathbb{P}_2) \rightarrow 0 \iff m = o(n).$$

Define $d := \prod_{i=1}^{m-1} \left(1 - \frac{i}{\binom{n}{2}}\right)$. Let μ denote the expected number of double edges. We have $\mu = \frac{m(m-1)}{\binom{n}{2}}$, and so for all $n \geq 2$ and $m \geq 1$, we have

$$|\mathcal{M}(n, m)|e^{-\mu}(1 - e^{\mu}(1 - d)) \leq |\mathcal{G}(n, m)| \leq |\mathcal{M}(n, m)|e^{-\mu}(1 + e^{\mu}(1 - d)).$$

Now we generalize to $\mathcal{G}(n, k_S, k_T)$ and $\mathcal{M}(n, k_S, k_T)$. Similar to the previous example, there exists a coupling between $\mathbb{P}_{\mathcal{G}}$ and $\mathbb{P}_{\mathcal{M}}$ so that the random graph model either generates the same graph in both $\mathcal{G}(n, k_S, k_T)$ and $\mathcal{M}(n, k_S, k_T)$, or generates a graph in $\mathcal{M}(n, k_S, k_T) \setminus \mathcal{G}(n, k_S, k_T)$. We have

$$|\mathcal{M}(n, k_S, k_T)| = \frac{\binom{n}{2}^{k_S}}{k_S!} \frac{\binom{n}{3}^{k_T}}{k_T!}.$$

Using Lemma 2.12, we estimate $|\mathcal{G}(n, k_S, k_T)|$ below.

Proposition 3.1. *Let λ and d_0 be defined as in Lemma 2.12. Then for all $n \geq 3$, $k_S \geq 0$, $k_T \geq 0$, we have*

$$\frac{\binom{n}{2}^{k_S}}{k_S!} \frac{\binom{n}{3}^{k_T}}{k_T!} e^{-\lambda}(1 - e^{\lambda}d_0) \leq |\mathcal{G}(n, k_S, k_T)| \leq \frac{\binom{n}{2}^{k_S}}{k_S!} \frac{\binom{n}{3}^{k_T}}{k_T!} e^{-\lambda}(1 + e^{\lambda}d_0). \quad (10)$$

For $k_S = O(n)$ and $k_T = O(n)$, asymptotically as $n \rightarrow \infty$, we have

$$|\mathcal{G}(n, k_S, k_T)| \sim \frac{\binom{n}{2}^{k_S}}{k_S!} \frac{\binom{n}{3}^{k_T}}{k_T!} e^{-\lambda}.$$

Proof. Simply rearrange Equation (9) and note that λ stays bounded if and only if $k_S = O(n)$ and $k_T = O(n)$ as $n \rightarrow \infty$, and $d_0 \rightarrow 0$ for these values of parameters. \square

3.2 A confidence interval for $C(G)$

With a bound on the total variation distance between $\mathbb{P}_{\mathcal{G}}$ and $\mathbb{P}_{\mathcal{M}}$, we can compute a confidence interval for $C(G)$ in $\mathcal{G}(n, k_S, k_T)$. Hence, define $C : \Omega_{\mathcal{M}} \rightarrow [0, 1]$, to be the random variable which specifies the clustering coefficients.

For each pair of distinct edges a, b , where $1 \leq a < b \leq k_S$, we let

$$\hat{X}_{a,b} = 1(\text{edges } a \text{ and } b \text{ share exactly one node}).$$

The total number of connected triplets, which we denote by W , is given by $W = \sum_{a,b} \hat{X}_{a,b}$. In fact, the collection of random variables $\{\hat{X}_{a,b}\}_{1 \leq a < b \leq k_S}$ is an i.i.d. sequence! Thus, W is *exactly* binomial with parameters $n = \binom{k_S}{2}$ and $p = \frac{n-2}{\binom{n}{2}}$, hence

$$\mathbb{E} W = \binom{k_S}{2} \frac{n-2}{\binom{n}{2}}$$

and

$$\text{Var}(W) = \binom{k_S}{2} \frac{n-2}{\binom{n}{2}} \left(1 - \frac{n-2}{\binom{n}{2}}\right).$$

Let $\lambda = \mathbb{E} W$ and $\sigma = \sqrt{\text{Var}(W)}$. For $k_S = O(\sqrt{n})$, W is asymptotically Poisson distributed with parameter λ , and for $k_S/\sqrt{n} \rightarrow \infty$, W is asymptotically normally distributed mean λ and variance σ^2 .

In terms of the clustering coefficient C , we have

$$C = \frac{3k_T}{3k_T + W}.$$

A $(1 - \alpha)$ level 2-sided confidence interval for C is given by any numbers $L(C)$ and $U(C)$ that satisfy

$$\mathbb{P}_{\mathcal{G}}(L(C) \leq C \leq U(C)) \geq 1 - \alpha.$$

Rearranging, we have

$$\mathbb{P}_{\mathcal{G}}\left(3k_T \left(\frac{1}{U(C)} - 1\right) \leq W \leq 3k_T \left(\frac{1}{L(C)} - 1\right)\right) \geq 1 - \alpha.$$

Let us consider the case when W is asymptotically normally distributed, i.e., $k_S/\sqrt{n} \rightarrow \infty$. Letting F and Φ denote the distribution functions of $(W - \lambda)$ and a normal random variable with mean 0 and variance σ^2 , respectively, the Berry–Esseen theorem, as improved in [16], is given by

$$\sup_x |F(x) - \Phi(x)| \leq \frac{0.33554(\rho + 0.415\sigma^3)}{\sigma^3\sqrt{n}},$$

where

$$\rho = \mathbb{E} |\hat{X}_{a,b} - \mathbb{E} \hat{X}_{a,b}|^3 = \frac{n-2}{\binom{n}{2}} \left(1 - \frac{n-2}{\binom{n}{2}}\right) \left(\left(1 - \frac{n-2}{\binom{n}{2}}\right)^2 + \left(\frac{n-2}{\binom{n}{2}}\right)^2\right).$$

To form a confidence interval for $C(G)$ under measure $\mathbb{P}_{\mathcal{G}}$, we now work backwards from W under measure $\mathbb{P}_{\mathcal{M}}$. First, we find a $(1 - \beta)$ level 2-sided confidence interval for a random variable Z from the standard normal distribution, where

$$\beta = \max(0, \alpha - \sup_x |F(x) - \Phi(x)| - d_{TV}(\mathbb{P}_{\mathcal{G}}, \mathbb{P}_{\mathcal{M}})).$$

Call the lower and upper bounds L_{β} and U_{β} , respectively. Then we replace random variable Z with random variable $(W - \lambda)/\sigma$, and rearrange to obtain

$$\mathbb{P}_{\mathcal{M}}(\sigma L_{\beta} + \lambda \leq W \leq \sigma U_{\beta} + \lambda) \geq 1 - \alpha,$$

whence,

$$L(C) = \left(\frac{\sigma U_{\beta} + \lambda}{3k_T} + 1\right)^{-1}, \quad U(C) = \left(\frac{\sigma L_{\beta} + \lambda}{3k_T} + 1\right)^{-1},$$

i.e., $[L(C), U(C)]$ is a $(1 - \alpha)$ level confidence interval for C in $\mathbb{P}_{\mathcal{G}}$.

When $\lambda = O(1)$, W is asymptotically Poisson distributed with parameter λ , and instead of a 2-sided confidence interval, we compute a 1-sided confidence interval. A $(1 - \alpha)$ level 1-sided confidence interval for C in this case is given by any number $V(C)$ such that

$$\mathbb{P}_{\mathcal{G}}(C \geq V(C)) \leq \alpha.$$

Rearranging, we see that the equivalent formulation in terms of W is

$$\mathbb{P}_{\mathcal{M}}\left(W \leq \lambda + 3k_T \left(\frac{1}{\sigma V(C)} - 1\right)\right) \leq \alpha.$$

Suppose k_α is the largest integer value such that

$$\mathbb{P}_{\mathcal{M}}(W \leq k_\alpha) \leq \alpha.$$

Then we let $\beta = \max(0, \alpha - d_{TV}(\mathbb{P}_{\mathcal{G}}, \mathbb{P}_{\mathcal{M}}))$, and appeal to a table of Binomial probabilities¹ and rearrange, to obtain

$$V(C) = \frac{1}{\sigma} \left(\frac{k_\beta + \lambda}{3k_T} + 1 \right)^{-1}.$$

References

- [1] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, pages 9–25, 1989.
- [2] Shweta Bansal, Shashank Khandelwal, and Lauren Ancel Meyers. Evolving clustered random networks. *arXiv preprint arXiv:0808.0509*, 2008.
- [3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [4] Andrew D Barbour, Lars Holst, and Svante Janson. *Poisson approximation*. Clarendon Press Oxford, 1992.
- [5] Mindaugas Bloznelis. Degree and clustering coefficient in sparse random intersection graphs. *The Annals of Applied Probability*, 23(3):1254–1289, 2013.
- [6] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.
- [7] Béla Bollobás and Oliver M Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.
- [8] Louis HY Chen. Poisson approximation for dependent trials. *The Annals of Probability*, pages 534–545, 1975.

¹We can alternatively use LeCam’s approximation, which states that $d_{TV}(\mathcal{L}(W), \text{Poisson}(\lambda)) \leq np^2$, but the resulting inversion of probabilities still requires a table of values.

- [9] Jesper Dall and Michael Christensen. Random geometric graphs. *Physical Review E*, 66(1):016121, 2002.
- [10] Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.
- [11] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, pages 1141–1144, 1959.
- [12] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical review E*, 65(2):026107, 2002.
- [13] Konstantin Klemm and Víctor M Eguíluz. Highly clustered scale-free networks. *Physical Review E*, 65(3):036123, 2002.
- [14] Mark E J Newman. Random graphs with clustering. *Physical review letters*, 103(5):058701, 2009.
- [15] M Angeles Serrano and Marián Boguná. Tuning clustering in random networks with arbitrary degree distributions. *Physical Review E*, 72(3):036133, 2005.
- [16] Irina Shevtsova. On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.
- [17] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.
- [18] Osman Yagan and Armand M Makowski. Random key graphs can they be small worlds? In *Networks and Communications, 2009. NETCOM’09. First International Conference on*, pages 313–318. IEEE, 2009.